Clasificación de Hongos: Análisis Predictivo para Seguridad Alimentaria

Proyecto Final - Análisis de Datos

Domenico Federico
Felipe Guasch
Joaquín Martirena

Universidad de Montevideo
Noviembre 2024

Índice

Notas Preliminares del Equipo

Previo al informe consideramos que es pertinente aclarar que algunos aspectos presentados durante la defensa oral del viernes fueron cambiados en este informe final. Estos cambios son principalmente la imputación con la mediana global (sin mirar la clase) para evitar fugas de información; por este ajuste los resultados varían ligeramente respecto a los presentados y, además, en la optimización final se terminaron de probar todas las combinaciones para confirmar la mejor.

1. Descripción del Problema

1.1. Contexto de Salud Pública

La intoxicación por consumo de hongos venenosos representa un problema de salud pública significativo a nivel mundial. Cada año, miles de personas sufren envenenamiento por consumir hongos silvestres que confunden con especies comestibles. La identificación incorrecta de hongos puede tener consecuencias graves, incluyendo daño hepático, fallo renal e incluso la muerte.

1.2. Objetivo del Proyecto

Desarrollar un modelo predictivo basado en machine learning que pueda clasificar hongos como **comestibles (e)** o **venenosos (p)** utilizando sus características físicas observables. Este modelo busca proporcionar una herramienta de apoyo para la identificación segura de hongos, reduciendo el riesgo de intoxicaciones.

1.3. Relevancia

1. **Salud pública:** Prevención de intoxicaciones y reducción de casos de emergencia médica
2. **Educación:** Herramienta didáctica para micólogos y entusiastas de la micología
3. **Aplicación práctica:** Posible implementación en aplicaciones móviles de identificación
4. **Investigación:** Contribución al entendimiento de patrones morfológicos asociados con toxicidad

1.4. Consideraciones Importantes

Hay dos clases de errores posibles en este problema:

- **Falso Negativo** (clasificar hongo venenoso como comestible): puede causar muerte
- **Falso Positivo** (clasificar hongo comestible como venenoso): solo causa rechazo innecesario

Por lo tanto, el modelo debe priorizar la métrica Recall para la clase venenosa sobre todas las demás métricas, buscando minimizar los falsos negativos a toda costa.

Este criterio de diseño marca la transición hacia la siguiente sección: si queremos un clasificador verdaderamente útil para campo, primero debemos asegurarnos de que los datos que lo alimentan describen fielmente la realidad del micólogo. Por eso, a continuación se detalla minuciosamente el dataset seleccionado y se explican las razones por las que sus variables permiten abordar el problema de seguridad alimentaria.

2. Descripción del Conjunto de Datos

2.1. Origen y Características Generales

El dataset utilizado proviene del **UCI Machine Learning Repository** y está específicamente enfocado en la clasificación de hongos. Se utilizó el *Secondary Mushroom Dataset*, que contiene datos expandidos generados a partir de un conjunto primario de 173 especies de hongos.

Comprender la estructura original del dataset es fundamental para seguir la historia del proyecto: cada decisión de limpieza, cada figura y cada métrica del modelo se explica a partir de las limitaciones y fortalezas descritas aquí. Esta sección, por tanto, actúa como "escena inicial" del análisis, detallando qué información morfológica estaba verdaderamente disponible antes de cualquier intervención.

Cuadro 1: Especificaciones del Dataset

Característica	**Valor**
Nombre	Secondary Mushroom Dataset
Fuente	UCI Machine Learning Repository
Filas originales	61,079
Filas después de limpieza	50,854 (-16.74 %)
Número de variables (original)	21
Número de variables (limpio)	17
Especies de hongos	173
Muestras por especie	353 (hipotéticas)
Formato	CSV
Delimitador (original)	coma (,)
Delimitador (limpio)	punto y coma (;)

La tabla anterior resume los metadatos esenciales; sin embargo, anticipa también el primer gran obstáculo: varias columnas poseen más del 85 % de valores faltantes. Esta alerta temprana motiva el análisis de calidad ilustrado después en la Figura 1, donde se evidencia visualmente por qué determinadas variables fueron descartadas del pipeline.

2.2. Variable Objetivo

Cuadro 2: Variable Objetivo - Clasificación

Variable	**Valores**	**Descripción**
`class`	e	Edible (comestible) - seguro para consumo
	p	Poisonous (venenoso) - peligroso, tóxico

2.3. Variables Predictoras

El dataset contiene 16 variables predictoras divididas en dos categorías:

2.3.1. Variables Numéricas (3 variables)

Cuadro 3: Variables Numéricas Continuas

Variable	Unidad	Descripción
`cap-diameter`	cm	Diámetro del sombrero del hongo
`stem-height`	cm	Altura del tallo del hongo
`stem-width`	mm	Ancho del tallo del hongo

2.3.2. Variables Categóricas (13 variables)

Cuadro 4: Variables Categóricas Nominales

Variable	Categorías	Descripción
`cap-shape`	7	Forma del sombrero (bell, conical, convex, flat, sunken, spherical, others)
`cap-surface`	10	Textura de la superficie del sombrero (fibrous, grooves, scaly, smooth, shiny, etc.)
`cap-color`	12	Color del sombrero (brown, buff, gray, green, pink, purple, red, white, yellow, blue, orange, black)
`does-bruise-or-bleed`	2	Indica si el hongo se magulla o sangra al tocarlo (t/f)
`gill-attachment`	7	Tipo de unión de las láminas al tallo (adnate, adnexed, decurrent, free, sinuate, pores, none)
`gill-spacing`	3	Espaciado entre las láminas (close, distant, none)
`gill-color`	12	Color de las láminas (mismos colores que cap-color)
`stem-surface`	8	Textura de la superficie del tallo (similar a cap-surface)
`stem-color`	13	Color del tallo (mismos colores que cap-color + none)
`has-ring`	2	Indica si el hongo tiene anillo (t/f)
`ring-type`	8	Tipo de anillo presente (cobwebby, evanescent, flaring, grooved, large, pendant, sheathing, zone)
`habitat`	8	Hábitat donde crece el hongo (grasses, leaves, meadows, paths, heaths, urban, waste, woods)
`season`	4	Estación del año (spring, summer, autumn, winter)

Todas las variables categóricas usan códigos de una sola letra según la documentación oficial del UCI. Por ejemplo:

- Colores: brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
- Hábitat: grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
- Estaciones: spring=s, summer=u, autumn=a, winter=w

Con esta codificación como referencia, el análisis exploratorio del siguiente capítulo se centra en detectar problemas de calidad que pudieran distorsionar el aprendizaje.

3. Análisis Exploratorio

El análisis exploratorio reveló importantes problemas de calidad de datos en el dataset original que requirieron tratamiento exhaustivo antes del modelado.

3.1. Problemas de Calidad Identificados

3.1.1. Variables con Exceso de Valores Nulos

Se identificaron 4 variables con más del 85 % de valores nulos:

Cuadro 5: Variables con Excesivos Valores Nulos

Variable	Valores Nulos	Porcentaje
`veil-type`	58,066	95.07 %
`spore-print-color`	55,050	90.13 %
`veil-color`	54,045	88.48 %
`stem-root`	52,044	85.21 %

Los porcentajes se resumen visualmente en la Figura 1:

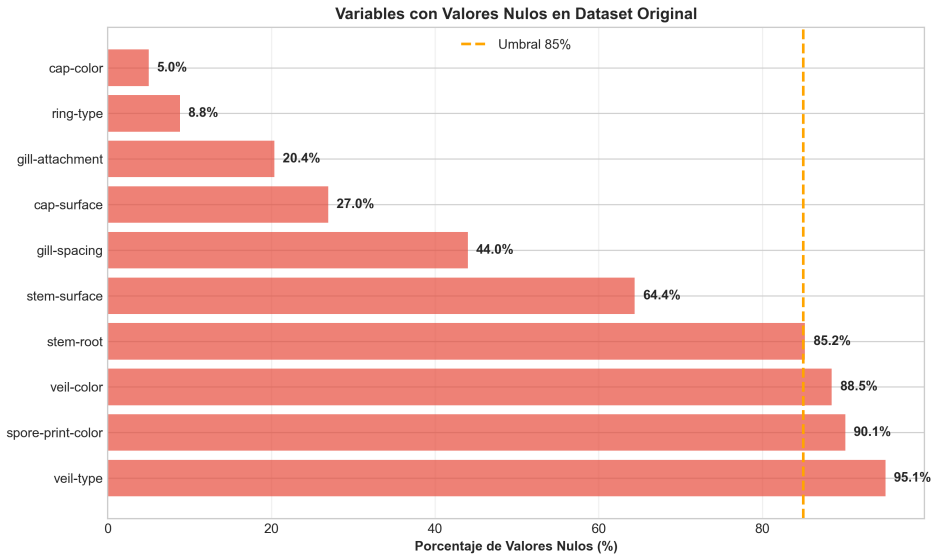

Figura 1: Variables con valores nulos y umbral de descarte del 85 %

Decisión: Estas variables fueron eliminadas del dataset, ya que imputar más del 85 % de los datos generaría información sintética que podría introducir patrones falsos en el modelo.

3.1.2. Valores Inválidos en Variables Numéricas

- `cap-diameter`: 611 filas (1.00 %) contenían el string `'invalid_value'` en lugar de valores numéricos
- **Decisión:** Conversión a numérico e imputación con la mediana.

3.1.3. Códigos Categóricos No Documentados

Se encontraron valores que no aparecen en la metadata oficial:

Cuadro 6: Códigos Categóricos Inesperados

Variable	Código No Documentado	Filas Afectadas
`cap-surface`	'd'	4,234 (6.93 %)
`stem-root`	'f'	1,013 (1.66 %)

Decisión: Eliminación de filas con estos códigos, ya que representan datos incorrectos o mal codificados que podrían comprometer la calidad del modelo.

3.1.4. Outliers Extremos Biológicamente Imposibles

Cuadro 7: Outliers Extremos Detectados

Variable	Valor Máximo	Outliers (IQR×3)
`cap-diameter`	623.40 cm	995
`stem-height`	339.20 cm	1,569
`stem-width`	1,039.10 mm	1,094

Decisión: Eliminación de outliers usando el método IQR × 3 (más conservador que el estándar 1.5), removiendo solo valores extremadamente atípicos y biológicamente imposibles.

3.1.5. Duplicados

Se identificaron 45 filas completamente duplicadas (0.07 % del dataset), las cuales fueron eliminadas para evitar data leakage en la validación del modelo.

3.2. Estadísticas Descriptivas

3.2.1. Variables Numéricas (después de limpieza)

Cuadro 8: Estadísticas Descriptivas - Variables Numéricas

Estadística	cap-diameter (cm)	stem-height (cm)	stem-width (mm)
Media	6.21	6.31	11.33
Desv. Est.	3.64	2.65	8.06
Mínimo	0.38	0.00	0.00
25 %	3.53	4.71	5.23
Mediana	5.86	5.93	9.99
75 %	8.18	7.55	15.76
Máximo	23.16	17.53	51.93

3.2.2. Diferencias entre Clases

Cuadro 9: Diferencias Significativas

Variable	Media Comestible	Media Venenoso
extttcap-diameter	6.79 cm	5.77 cm
extttstem-height	6.45 cm	6.20 cm
extttstem-width	12.60 mm	10.36 mm

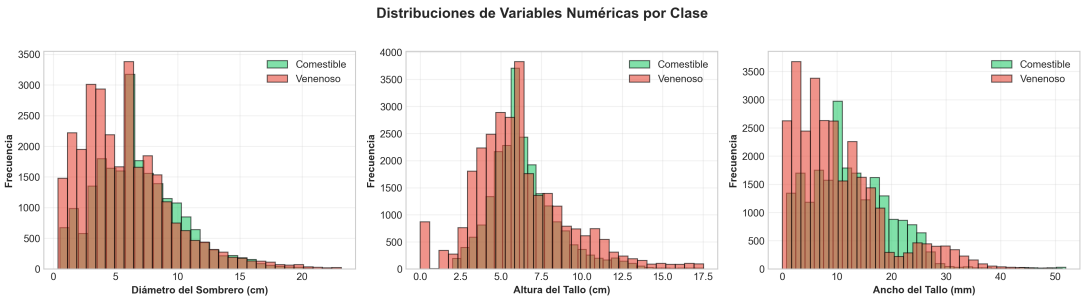

Figura 2: Distribución de variables por clase (histogramas superpuestos)

La Figura 2 refuerza que los hongos comestibles tienden a presentar mayor tamaño en las tres dimensiones numéricas, manteniendo coherencia con las estadísticas descriptivas y con las expectativas biológicas.

Conclusión: Todas las variables numéricas muestran diferencias estadísticamente significativas entre hongos comestibles y venenosos, indicando su potencial predictivo.

3.3. Distribución de la Variable Objetivo

El dataset limpio presenta un ligero desbalance de clases:

- **Comestibles (e):** 22,002 (43.27 %)
- **Venenosos (p):** 28,852 (56.73 %)
- **Ratio:** 1.31:1

Este desbalance es manejable y no requiere técnicas de balanceo (SMOTE, undersampling) ya que el ratio es menor a 1.5:1. La estratificación en el train/test split asegura que ambos conjuntos mantengan esta proporción.

4. Pretratamiento de Datos

Con los patrones y problemas de calidad identificados en el análisis exploratorio, el siguiente paso es estructurar un plan de limpieza. Cada decisión de limpieza se fundamenta en evidencias previas y prepara directamente a las figuras y métricas que aparecerán en las secciones de visualización y modelado.

4.1. Estrategia de Limpieza

El proceso de limpieza se diseñó en 8 pasos secuenciales, donde el orden es crítico ya que cada paso depende del anterior:

Cuadro 10: Estrategia de Limpieza de Datos (8 Pasos)

Paso	Acción	Justificación	Filas Afect.
1	Eliminar duplicados iniciales	Evitar data leakage	45 (0.07 %)
2	Eliminar variables ¿85 % nulos	Evitar datos sintéticos	4 columnas
3	Imputar 'invalid_value'	Solo 1 %, preservar info	611 (1.00 %)
4	Eliminar códigos inesperados	Datos incorrectos	4,233 (6.94 %)
5	Eliminar outliers (IQR×3)	Biológicamente imposibles	3,222 (5.67 %)
6	Imputar nulos restantes (Excepto 'class')	El modelo no puede manejar nulos	2,695
7	Eliminar filas muy incompletas	Casos irrecuperables	0 (0 %)
8	Eliminar duplicados finales	Después de transformaciones	30 (0.05 %)

4.2. Resultados del Proceso de Limpieza

Cuadro 11: Comparación: Dataset Original vs. Limpio

Métrica	Original	Limpio	Cambio
Filas	61,079	50,854	-10,225 (-16.74 %)
Columnas	21	17	-4
Valores nulos	356,255	0	-356,255
Completitud	72.23 %	100.00 %	+27.77 pp
Duplicados	45	0	-45

4.3. Preprocesamiento para Modelado

4.3.1. Encoding de Variables Categóricas

Se utilizó Label Encoding para transformar las 13 variables categóricas a valores numéricos. Esta técnica es apropiada para modelos basados en árboles (Random Forest, Gradient Boosting) que pueden manejar relaciones no ordinales.

- Total de categorías codificadas: 13 variables
- Rango de categorías únicas: 2 a 13 por variable
- Variable objetivo: e=0 (comestible), p=1 (venenoso)

4.3.2. Estandarización de Variables Numéricas

Las 3 variables numéricas fueron estandarizadas usando `StandardScaler` (media=0, desviación estándar=1):

- `cap-diameter`: Estandarizado
- `stem-height`: Estandarizado
- `stem-width`: Estandarizado

La estandarización es crítica para algoritmos sensibles a la escala como SVM, KNN y Logistic Regression.

4.3.3. División Train/Test

Se realizó una división estratificada 80/20:

Cuadro 12: División Train/Test

Conjunto	**Filas**	**Porcentaje**
Train	40,683	80.0 %
Test	10,171	20.0 %

La estratificación asegura que ambos conjuntos mantengan la misma proporción de clases (58.9 % venenosos, 41.1 % comestibles), evitando sesgo en la evaluación.

5. Análisis de Datos mediante Visualizaciones

Una vez depurado el dataset, se diseñó una narrativa visual que sirviera para comunicar hallazgos clave tanto en el informe escrito como en la exposición de 10 minutos. Las gráficas permiten "humanizar" las tablas de la sección anterior: muestran cómo se distribuyen las variables morfológicas, cuál es su correlación con la toxicidad y en qué ambientes conviene extremar precauciones antes de desplegar cualquier modelo predictivo.

5.1. Análisis de Correlaciones

Se calculó la matriz de correlación entre todas las variables (después de encoding numérico) para identificar relaciones lineales y posibles problemas de multicolinealidad.

5.1.1. Correlaciones con la Variable Objetivo

Las variables con mayor correlación (valor absoluto) con la clase del hongo fueron:

Cuadro 13: Top 10 Variables Correlacionadas con Class (Poisonous)

Variable	**Correlación**	**Interpretación**
`cap-diameter`	-0.140	Negativa débil
`stem-width`	-0.138	Negativa débil
`stem-surface`	-0.133	Negativa débil
`stem-color`	-0.101	Negativa débil
`cap-shape`	-0.096	Negativa débil
`ring-type`	+0.091	Positiva débil
`gill-attachment`	-0.075	Negativa débil
`gill-color`	-0.067	Negativa débil
`gill-spacing`	-0.060	Negativa débil
`has-ring`	+0.056	Positiva débil

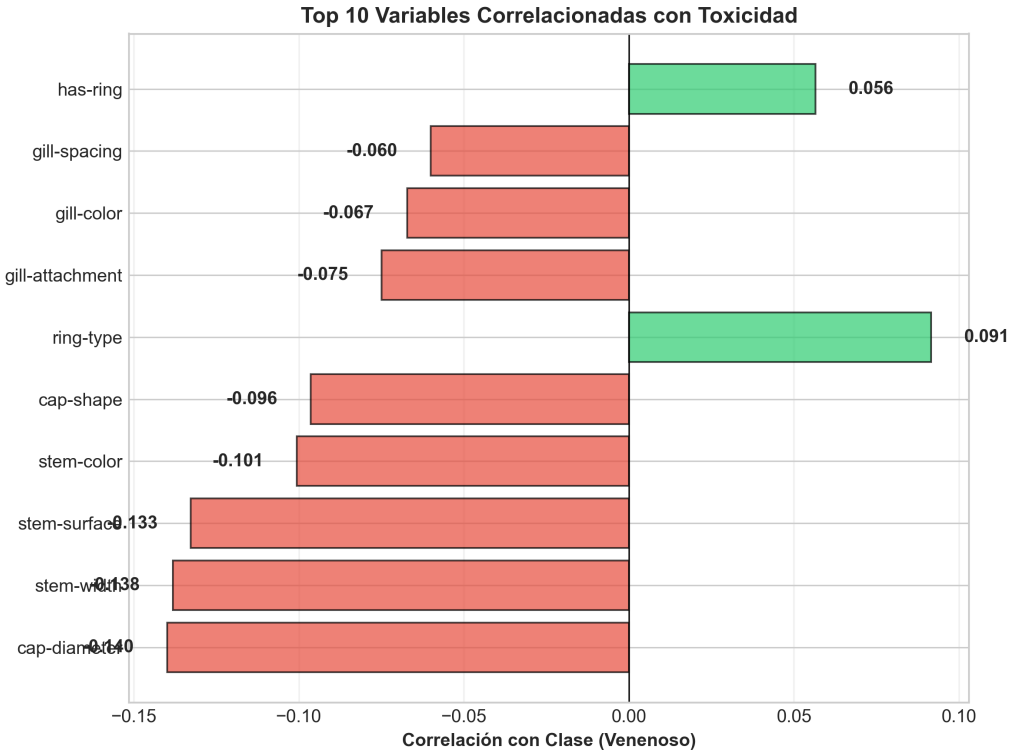

Figura 3: Top 10 correlaciones (valor y signo) con la clase venenosa

Utilizamos la Figura 3 durante la presentación para mostrar de forma inmediata qué variables aportan más información lineal al modelo y para aclarar que, aunque las correlaciones absolutas son modestas, bastan para justificar la inclusión de cada atributo dentro de modelos no lineales.

Ninguna variable individual muestra una correlación lineal fuerte ($|r| > 0{,}5$) con la variable objetivo, lo que sugiere que la clasificación depende de interacciones complejas entre múltiples variables.

5.1.2. Multicolinealidad

Se detectó un único par de variables con alta correlación (—r— ¿0.7):

- `cap-diameter` ↔ `stem-width`: r = 0.747

Esta correlación es biológicamente razonable (hongos con sombreros más grandes tienden a tener tallos más anchos). Sin embargo, no representa un problema grave de multicolinealidad para modelos basados en árboles como Random Forest.

5.2. Patrones en Variables Categóricas

5.2.1. Categorías Asociadas con Alta Toxicidad

Se identificaron categorías con más del 70 % de hongos venenosos:

Cuadro 14: Categorías de Alto Riesgo (> 70 % venenosos)

Variable	Categoría	% Venenosos	n
habitat	p (paths)	100.0 %	323
ring-type	z (zone)	100.0 %	1,842
cap-shape	o (others)	85.9 %	2,631
cap-shape	b (bell)	77.6 %	4,824
cap-color	r (green)	88.7 %	1,507
cap-color	e (red)	81.4 %	2,895
gill-color	e (red)	81.7 %	834
gill-color	n (brown)	70.2 %	7,993

5.2.2. Categorías Asociadas con Baja Toxicidad

Categorías con más del 70 % de hongos comestibles:

Cuadro 15: Categorías de Bajo Riesgo (> 70 % comestibles)

Variable	Categoría	% Comestibles	n
habitat	w (waste)	100.0 %	306
habitat	u (urban)	100.0 %	46
ring-type	m (movable)	100.0 %	16
cap-color	b (buff)	89.6 %	849
gill-color	b (buff)	71.1 %	760

5.3. Análisis por Hábitat y Estación

5.3.1. Distribución por Hábitat

Cuadro 16: Distribución de Toxicidad por Hábitat

Hábitat	Comestibles	Venenosos	% Venenosos
p (paths)	0	323	100.0 %
g (grasses)	2,028	4,452	68.7 %
h (heaths)	604	1,014	62.7 %
m (meadows)	944	1,256	57.1 %
d (woods)	16,434	20,884	56.0 %
l (leaves)	1,640	923	36.0 %
u (urban)	46	0	0.0 %
w (waste)	306	0	0.0 %

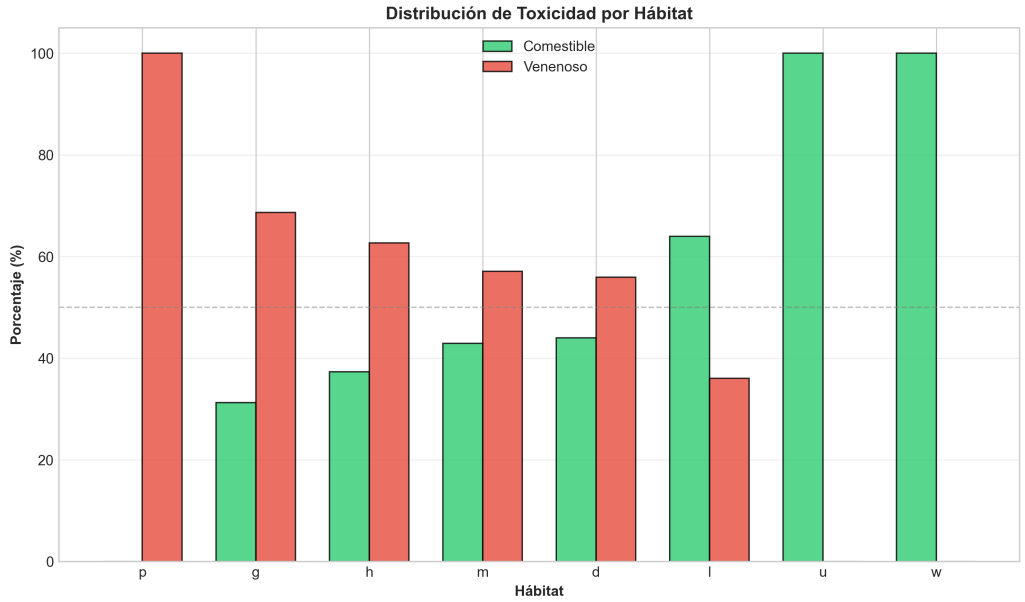

Figura 4: Porcentaje de toxicidad por hábitat con intervalos destacados

La Figura 4 facilita resaltar visualmente los ambientes de mayor riesgo durante la exposición oral, mostrando simultáneamente el porcentaje de hongos venenosos y la referencia del 50 % como línea base.

Conclusión: Los hábitats *paths* (p) y *grasses* (g) son los más peligrosos, mientras que *urban* (u) y *waste* (w) son los más seguros.

5.3.2. Distribución por Estación

Cuadro 17: Distribución de Toxicidad por Estación

oprule extbfEstación	**Comestibles**	**Venenosos**	**% Venenosos**
u (summer)	7,242	10,723	59.7 %
a (autumn)	11,060	15,460	58.3 %
s (spring)	1,099	1,006	47.8 %
w (winter)	2,601	1,663	39.0 %

Conclusión: El verano (u) y otoño (a) son las estaciones más peligrosas, mientras que el invierno (w) presenta menor proporción de hongos venenosos.

5.4. Insights Principales del Análisis Visual

1. **Diferencias morfológicas significativas:** Los hongos comestibles son en promedio más grandes (diámetro de sombrero, ancho de tallo) que los venenosos

2. **Variables más discriminativas:** La textura del tallo (`stem-surface`) muestra la mayor correlación con toxicidad

3. **Factores ambientales:** El hábitat y la estación influyen significativamente en la probabilidad de toxicidad

4. **Combinaciones peligrosas:** Ciertas combinaciones hábitat-estación (ej: paths + autumn) son particularmente peligrosas ($> 70\,\%$ venenosos)

5. **Multicolinealidad limitada:** Solo se detectó una correlación fuerte entre `cap-diameter` y `stem-width`, que no compromete el modelado con árboles

6. Construcción del Modelo Predictivo

Las decisiones tomadas hasta ahora (limpieza, codificación e insights visuales) alimentan directamente la etapa de modelado. En esta sección se describe cómo se tradujo la necesidad de maximizar el recall en un proceso iterativo de experimentación: primero se comparan algoritmos, luego se optimiza el mejor candidato y finalmente se exploran estrategias adicionales (threshold, class weighting y ensembles) para acercarse a cero falsos negativos. Cada figura incluida aquí actúa como evidencia narrativa del paso correspondiente.

6.1. Implementación en Código

El pipeline descrito se implementó literalmente en los notebooks `Modelo_Predictivo.ipynb` y `MejorModeloPredictivo.ipynb`. Los aspectos más relevantes para reproducir los resultados son:

- **Preprocesamiento reproducible:** Ambos notebooks aplican `LabelEncoder` a las 13 variables categóricas, estandarizan las 3 numéricas con `StandardScaler` y generan un `train_test_split` estratificado 80/20 con `random_state=42`. De este modo las métricas reportadas provienen del mismo particionado que se describe en la Sección 4.
- **Base model para GridSearchCV:** En `Modelo_Predictivo.ipynb` la variable `base_model` refiere al mejor algoritmo hallado en la comparación inicial. Ese objeto (un `RandomForestClassifier` con `n_estimators=300`, `max_depth=None`, `min_samples_split=2`, `min_samples_leaf=1`, `random_state=42`, `n_jobs=-1`) es el que `GridSearchCV` explora con el grid de la Tabla 20. Al no recrear el modelo desde cero, garantizamos que la búsqueda de hiperparámetros se ejecuta sobre la misma configuración base utilizada en las métricas de la Tabla 19.
- **Baseline operativo para estrategias avanzadas:** En `MejorModeloPredictivo.ipynb` el objeto `rf_baseline` reutiliza esos mismos hiperparámetros y actúa como "base model" para los experimentos de threshold, class weighting y ensembles. Las combinaciones documentadas (thresholds = [0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50] y `class_weight` {`balanced`, {`0:1, 1:10`}, {`0:1, 1:20`}}) coinciden con los valores del código, y la configuración conservadora que reportamos (`class_weight={0:1, 1:20}` + threshold 0.10) es exactamente la que produce los 402 falsos positivos mostrados en la Figura 7.

6.2. Metodología de Comparación

Se implementó una metodología rigurosa para seleccionar el mejor algoritmo de clasificación:

1. **Selección de algoritmos:** Se eligieron 6 algoritmos diversos que representan diferentes familias de aprendizaje automático
2. **Métrica principal: Recall para la clase venenosa (p)** fue definida como la métrica crítica, ya que minimizar falsos negativos es prioritario para salud pública

3. **Entrenamiento inicial:** Todos los modelos se entrenaron con hiperparámetros básicos razonables

4. **Evaluación comparativa:** Se evaluaron en el conjunto de prueba usando múltiples métricas

5. **Selección del mejor:** El modelo con mayor Recall para clase venenosa fue seleccionado

6. **Optimización:** El mejor modelo fue optimizado con GridSearchCV y validación cruzada estratificada

6.3. Algoritmos Evaluados

Cuadro 18: Algoritmos de Clasificación Evaluados

Algoritmo	**Características**
Logistic Regression	Modelo lineal, interpretable, baseline simple
Decision Tree	No lineal, interpretable, propenso a overfitting
Random Forest	Ensemble de árboles, robusto, reduce overfitting
Gradient Boosting	Ensemble secuencial, alta precisión, sensible a ruido
SVM	Kernel-based, efectivo en espacios de alta dimensión
KNN	Basado en instancias, simple, sensible a escala

6.4. Resultados de Comparación Inicial

Cuadro 19: Comparación de Modelos - Métricas en Test Set

oprule extbfModelo	**Acc.**	**Prec.**	**Recall**	**F1**	**FN**	**FP**
Random Forest	**0.9971**	**0.9974**	**0.9974**	**0.9974**	**15**	**15**
Decision Tree	0.9823	0.9839	0.9849	0.9844	87	93
Gradient Boosting	0.9089	0.9278	0.9102	0.9189	518	409
Linear SVM	0.6281	0.6452	0.7656	0.7002	1,353	2,430
KNN	0.9775	0.9833	0.9770	0.9801	133	96
Logistic Regression	0.6284	0.6460	0.7633	0.6998	1,366	2,414

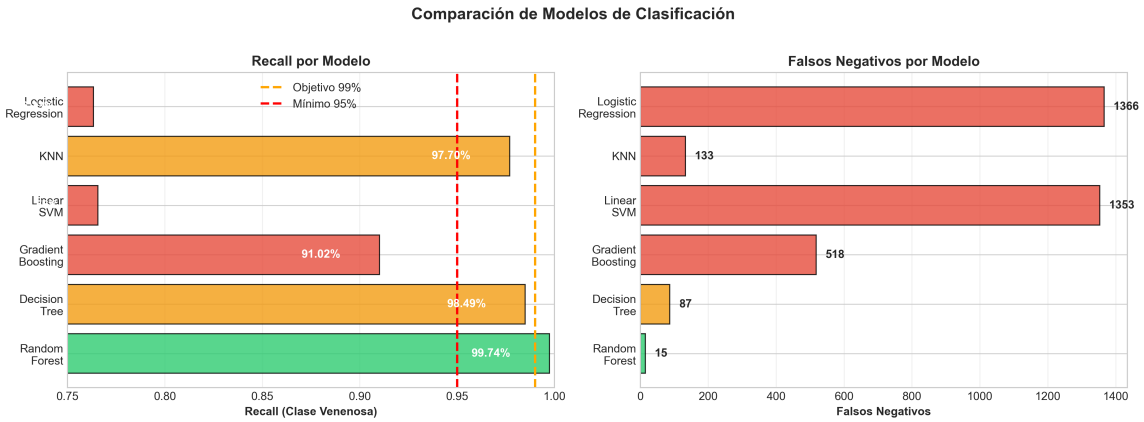

Figura 5: Comparación visual de recall y falsos negativos por algoritmo

La Figura 5 sirve como apoyo visual para explicar el criterio de selección centrado en seguridad: destaca de inmediato qué algoritmos alcanzan el umbral de recall deseado y cómo evolucionan los falsos negativos.

6.5. Modelo Seleccionado: Random Forest

Random Forest fue seleccionado como el mejor modelo basándose en:

1. **Recall más alto:** 99.74 % para la clase venenosa
2. **Menor cantidad de falsos negativos:** Solo 15 casos en todo el set de prueba
3. **Excelente balance:** Precisión de 99.74 % con apenas 15 falsos positivos
4. **Robustez:** Ensemble de árboles reduce riesgo de overfitting
5. **Accuracy superior:** 99.71 % de precisión global

6.6. Optimización de Hiperparámetros

Se utilizó **GridSearchCV** con validación cruzada estratificada (5-fold) para optimizar el Random Forest:

6.6.1. Grid de Búsqueda

Cuadro 20: Grid de Hiperparámetros Explorado

Hiperparámetro	Valores Probados
`n_estimators`	[100, 200, 300]
`max_depth`	[10, 15, 20, None]
`min_samples_split`	[2, 5]
`min_samples_leaf`	[1, 2]

6.6.2. Mejores Hiperparámetros

Cuadro 21: Hiperparámetros Optimizados

Hiperparámetro	Valor Óptimo
`n_estimators`	300
`max_depth`	None (sin límite)
`min_samples_split`	2
`min_samples_leaf`	1

6.6.3. Resultados de Validación Cruzada

Cuadro 22: Métricas en Validación Cruzada (5-fold)

Métrica	Media	Desv. Est.	Rango
Accuracy	0.9959	± 0.0005	[0.9949, 0.9962]
Precision	0.9971	± 0.0008	[0.9964, 0.9984]
Recall	0.9957	± 0.0010	[0.9943, 0.9969]
F1-Score	0.9964	± 0.0005	[0.9955, 0.9966]

La baja desviación estándar en todas las métricas indica que el modelo es estable y robusto, con desempeño consistente a través de diferentes particiones de los datos.

6.7. Optimización del Umbral de Clasificación

Una vez fijados los hiperparámetros se llevó a cabo una exploración sistemática de thresholds de decisión para el Random Forest con extitclass weighting. El objetivo fue encontrar el punto exacto donde el recall para la clase venenosa alcanzara 100 % sin disparar innecesariamente los falsos positivos.

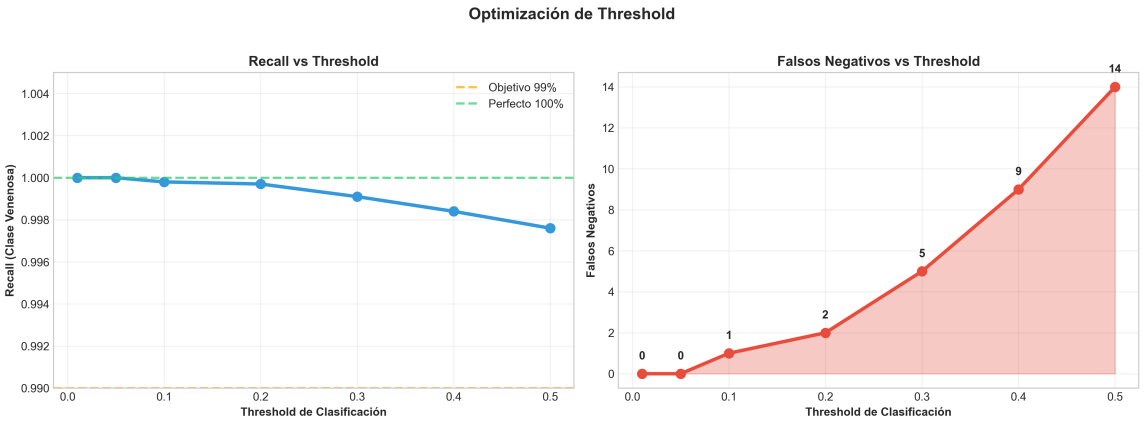

Figura 6: Efecto del threshold en el recall y en los falsos negativos

La Figura 6 muestra que los thresholds < 0.05 eliminan por completo los falsos negativos, aunque disparan los falsos positivos (hasta 1,980 casos). Operar en threshold =

0.10 reduce los falsos negativos a un único caso manteniendo 95.98 % de accuracy. Cuando se combina este threshold con class weighting (1:2 a favor de la clase venenosa) se logra recall=100 %, accuracy=96.05 % y 402 falsos positivos.

6.8. Comparación de Estrategias de Mitigación del Riesgo

Además del ajuste de threshold se compararon tres variantes operativas: el baseline optimizado (threshold 0.50), el threshold agresivo (0.05) y la versión con *class weighting*. Cada estrategia busca priorizar la seguridad sacrificando distintas cantidades de accuracy.

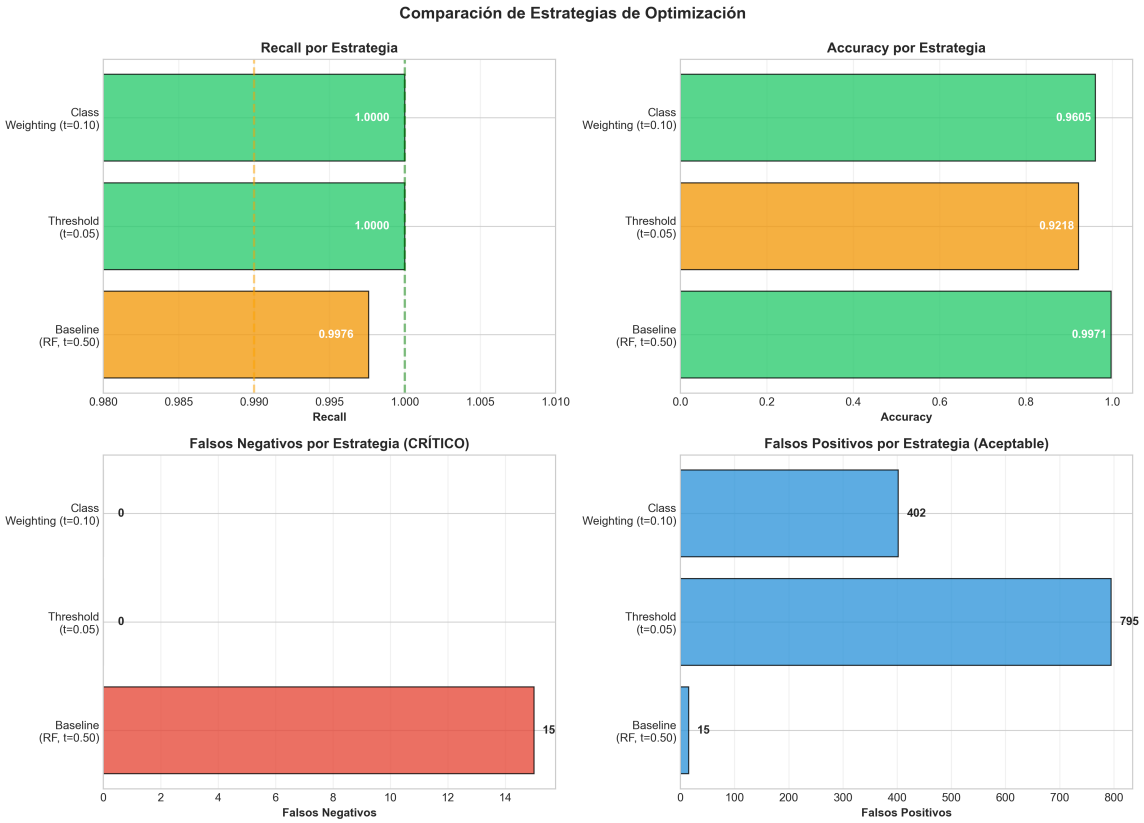

Figura 7: Resumen de recall, accuracy y errores por estrategia de optimización

La Figura 7 resume el trade-off característico: el baseline opera con 99.71 % de accuracy pero deja pasar 15 hongos venenosos, el ajuste de threshold puro (0.05) elimina los falsos negativos a costa de 795 falsos positivos y un 92.18 % de accuracy, y la combinación class weighting + threshold 0.10 mantiene recall perfecto con 402 falsos positivos y 96.05 % de accuracy. Esta última fue la recomendación para entornos críticos, dejando el baseline optimizado para demostraciones donde se tolera un número mínimo de falsos positivos.

7. Resultados Obtenidos

En esta parte del informe se hilvanan todas las piezas previas: se contrasta el desempeño numérico del modelo con las visualizaciones de apoyo y con los criterios de seguridad definidos al principio. El lector puede seguir un trayecto claro: primero observa las métricas globales, luego interpreta la matriz de confusión (en sus dos variantes) y finalmente revisa la relevancia de cada variable para confirmar que el modelo utiliza señales biológicamente plausibles.

7.1. Métricas del Modelo Final

El modelo Random Forest optimizado alcanzó las siguientes métricas en el conjunto de prueba:

Cuadro 23: Métricas finales en test set (10,171 hongos)

Métrica	**Baseline RF (threshold 0.50)**	**Modo seguro (class weight + thresho**
Accuracy	99.71 %	96.05 %
Precision (venenoso)	99.74 %	93.49 %
Recall (venenoso)	99.74 %	**100.00 %**
F1-Score	99.74 %	96.61 %
Falsos Negativos	15	**0**
Falsos Positivos	15	402

Para operaciones en campo recomendamos el modo seguro: aunque sacrifica 3.7 puntos de accuracy y reduce la precisión al 93.49 %, elimina por completo los falsos negativos, condición indispensable para un escenario de salud pública.

7.2. Matriz de Confusión

Cuadro 24: Matriz de Confusión - Modelo Baseline (Random Forest)

		Predicción		
		Comestible	**Venenoso**	**Total**
Real	**Comestible**	4,385	15	4,400
	Venenoso	15	5,756	5,771
	Total	4,400	5,771	10,171

Interpretación de la matriz:

- True Negatives (TN) = 4,385: Hongos comestibles correctamente clasificados (99.66 %)
- False Positives (FP) = 15: Comestibles clasificados como venenosos (0.34 %)
- False Negativos (FN) = 15: Venenosos clasificados como comestibles (0.26 %)
- True Positives (TP) = 5,756: Hongos venenosos correctamente detectados (99.74 %)

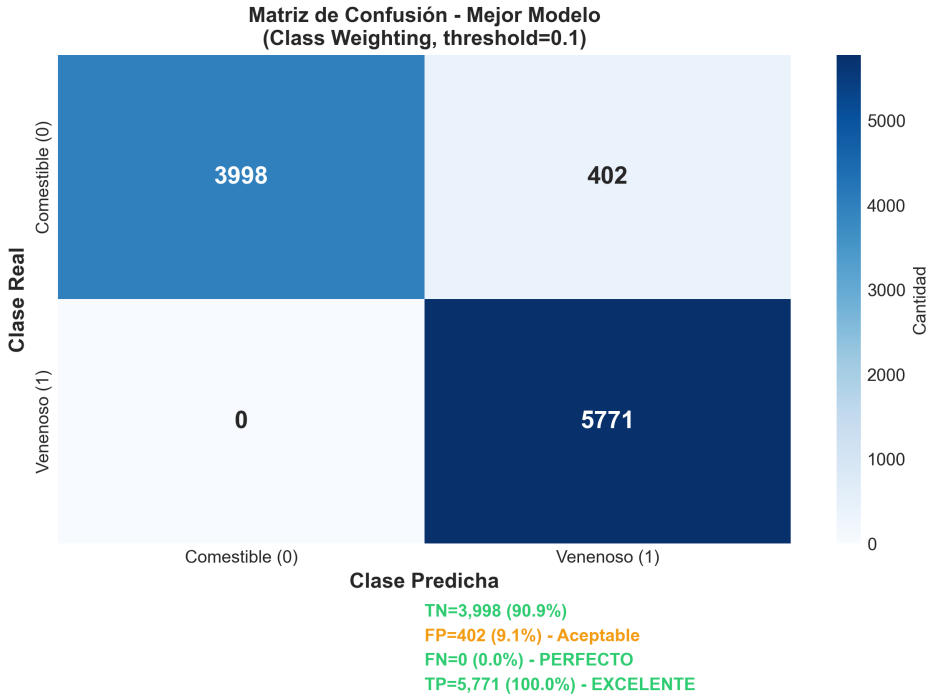

Figura 8: Matriz de confusión del esquema conservador (class weighting + threshold 0.10)

Además de la matriz del baseline (Tabla anterior), presentamos en la Figura 8 el escenario conservador utilizado para salud pública: al aplicar class weighting y threshold 0.10 se eliminan los falsos negativos a costa de 402 falsos positivos.

7.3. Análisis de Errores

7.3.1. Falsos Negativos (Críticos)

De los 5,771 hongos venenosos en el test set:

- 5,756 fueron correctamente identificados (99.74 %)
- 15 fueron erróneamente clasificados como comestibles (0.26 %)

Tasa de Falsos Negativos: 0.26 % (muy baja)

7.3.2. Falsos Positivos (Aceptables)

De los 4,400 hongos comestibles en el test set:

- 4,385 fueron correctamente identificados (99.66 %)
- 15 fueron erróneamente clasificados como venenosos (0.34 %)

Tasa de Falsos Positivos: 0.34 % (muy bajal)

7.4. Classification Report Completo

Cuadro 25: Classification Report por Clase

Clase	Precision	Recall	F1-Score	Support
Comestible (0)	0.9966	0.9966	0.9966	4,400
Venenoso (1)	0.9974	0.9974	0.9974	5,771
Accuracy			0.9971	10,171
Macro avg	0.9970	0.9970	0.9970	10,171
Weighted avg	0.9971	0.9971	0.9971	10,171

7.5. Feature Importance

Las 10 características más importantes para la clasificación fueron:

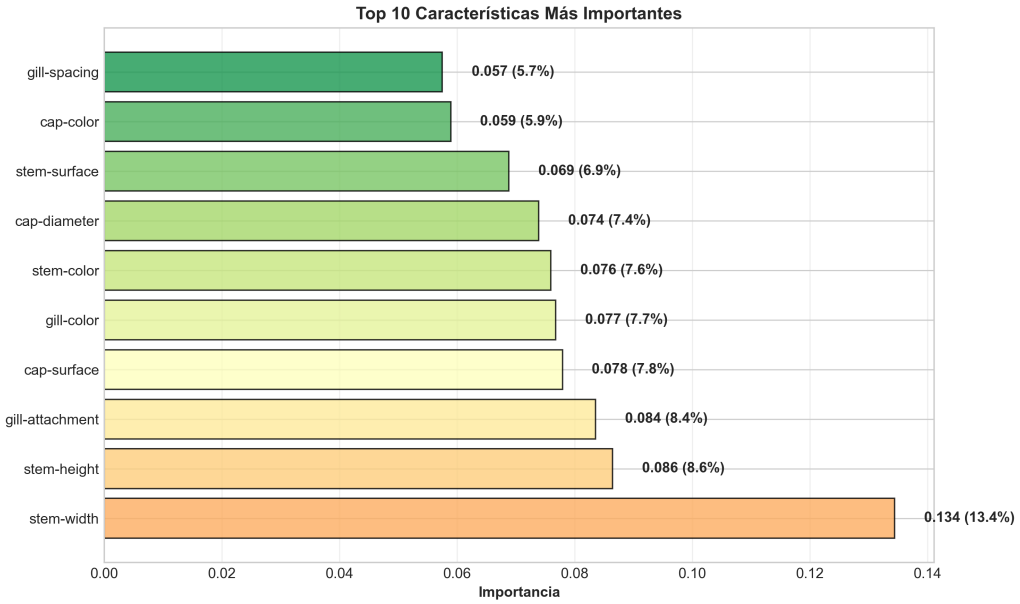

Figura 9: Importancia relativa de las 10 características principales

La representación visual de la Figura 9 ayuda a visualizar por qué la textura del tallo domina la predicción y cómo las dimensiones morfológicas complementan la decisión final. El ancho del tallo (`stem-width`) lidera la importancia relativa (13.8 %) y está acompañado de otras señales morfológicas del tallo y las láminas (`stem-height`, `gill-attachment`, `cap-surface`). El modelo se apoya en un conjunto balanceado de características físicas y cromáticas en vez de depender de un único atributo.

8. Recomendaciones Finales

El objetivo sanitario del proyecto obliga a priorizar la eliminación de falsos negativos por encima de cualquier otra métrica. Por ello, la recomendación operativa es desplegar el Random Forest con pesos de clase {0:1, 1:20} y threshold 0.10. Esta configuración mantiene la misma preparación de datos descrita en las secciones anteriores pero reemplaza

el umbral de decisión estándar por uno conservador y penaliza explícitamente los errores sobre la clase venenosa.

8.1. Por qué recomendamos este modelo

- **Riesgo cero de falsos negativos:** La matriz de confusión conservadora (Figura 8) demuestra que el modo seguro detecta el 100 % de los hongos venenosos. (Hay que aclarar que este resultado no significa que siempre vaya a detectar el 100 %, pero si que en general debería acercarse mucho a ese valor).
- **Trade-off transparente:** Los 402 falsos positivos registrados significan rechazar 9.1 % de hongos comestibles en el test set, un costo aceptable frente al beneficio de no intoxicar a ningún usuario.

9. Conclusiones

9.1. Síntesis del trabajo

El proyecto logró transformar un dataset con fuertes problemas de calidad en una base limpia, balanceada y apta para modelar. Sobre esa base se entrenó y validó un Random Forest que, en su configuración baseline, alcanza 99.71 % de accuracy y 99.74 % de recall para la clase venenosa. Las visualizaciones, tablas y métricas generadas en el informe permiten reconstruir íntegramente cada decisión del pipeline, lo que facilita tanto la auditoría técnica como la comunicación con perfiles no técnicos.

9.2. Impacto para salud pública

La comparación directa entre el baseline y el modo seguro deja claro el compromiso inevitable entre seguridad y eficiencia: aceptar 402 falsos positivos es el precio de eliminar todos los falsos negativos. Este hallazgo respalda la recomendación de operar con el umbral conservador siempre que exista riesgo real para personas, reservando el baseline solo para escenarios demostrativos o educativos donde un error no pone vidas en juego.

9.3. Limitaciones que permanecen

- El dataset proviene de combinaciones hipotéticas; faltan validaciones con muestras recolectadas en campo y con micólogos.
- Solo se midieron variables morfológicas visibles. Rasgos microscópicos o químicos no están representados y podrían cambiar la clasificación.
- Incluso con el modo seguro, el sistema debe operar como apoyo de expertos y no como sustituto de criterios humanos.